Δευτέρα 07/12/20

Υποθέσεις για τα σφάλματα

i) $E(\varepsilon_{ij}) = 0$, $i = 1, \dots I$, $j = 1, \dots J_i$

ii) $Var(\varepsilon_{ij}) = \sigma^2$

iii) $Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0$, $i \neq j$, $k \neq l$

iv) $\varepsilon_{ij} \sim N(0, \sigma^2)$

Συνέπειες:

i) $E(Y_{ij}) = \mu + a_i$

ii) $Var(Y_{ij}) = \sigma^2$

iii) Y_{ij} ασυσχέτιστα ανά δύο

iv) $Y_{ij} \sim N(\mu + a_i, \sigma^2)$

$\Big\}$ Y_{ij} ανεξάρτητα

Θα υποθέταμε ότι ικανοποιούνται

Θεώρημα:

Υπό τις υποθέσεις για τα σφάλματα ισχύει

α) $E(MS_{res}) = \sigma^2$

β) $E(MS_{tr}) = \sigma^2 + \dfrac{1}{I-1} \sum_{i=1}^{I} J_i a_i^2$

Απόδειξη

α) $MS_{res} = \frac{SS_{res}}{N-I} = \frac{1}{N-I} \sum_{i=1}^{I}\sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i.})^2$

Λόγω των υποθέσεων για τα σφάλματα και ειδικότερα από την 4η υπόθεση κάθε γραμμή δεδομένων προέρχεται από πληθυσμό με διακύμανση σ^2. Δηλαδή $Y_{i1}, Y_{i2}, \dots Y_{iJ_i}$ είναι τ.δ από πληθυσμό με διακύμανση σ^2.

Γνωρίζουμε ότι αν $W_1, \dots W_n$ τ.δ από πληθυσμό με διακύμανση σ^2 τότε $E(S_w^2) = \sigma^2$ όπου $S_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{W})^2$

Άρα $E(S^2_{Y_{iJ_i}}) = E\left(\frac{1}{J_i-1}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2\right) = \sigma^2$

$\Rightarrow E\left[\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2\right] = \sigma^2(J_i - 1)$. (1).

$E(MS_{res}) = E\left(\frac{1}{N-I}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2\right) =$

$= \frac{1}{N-I}\sum_{i=1}^{I} E\left[\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2\right] = \frac{1}{N-I}\sum(J_i - 1)\sigma^2 =$

$= \frac{\sigma^2}{N-I}\left(\sum_{i=1}^{I} J_i - I\right) = \frac{\sigma^2}{N-I}(N-I) = \sigma^2$

β) $MStr = \frac{SStr}{I-1} = \frac{1}{I-1}\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$

Από το μοντέλο $Y_{ij} = \mu + a_i + \varepsilon_{ij}$ και την πλευρική συνθήκη

$\sum_{i=1}^{I} J_i a_i = 0$ παίρνουμε ότι $Y_{i\cdot} = J_i\mu + J_i a_i + \varepsilon_{i\cdot}$

και $Y_{\cdot\cdot} = N\mu + \varepsilon_{\cdot\cdot}$

Έτσι $E(Y_{i\cdot}) = J_i(\mu + a_i)$, $Var(Y_{i\cdot}) = Var(\varepsilon_{i\cdot}) = J_i\sigma^2$

$E(Y_{\cdot\cdot}) = N\mu$, $Var(Y_{\cdot\cdot}) = N\sigma^2$

Επομένως

$$E(SStr) = E\left\{\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2\right\} = E\left\{\sum_{i=1}^{I} J_i\left(\frac{Y_{i\cdot}}{J_i} - \frac{Y_{\cdot\cdot}}{N}\right)^2\right\}$$

$$= E\left\{\sum_{i=1}^{I}\frac{Y_{i\cdot}^2}{J_i} - \frac{Y_{\cdot\cdot}^2}{N}\right\} = \sum_{i=1}^{I}\frac{1}{J_i}E(Y_{i\cdot}^2) - \frac{1}{N}E(Y_{\cdot\cdot}^2) =$$

$$= \sum_{i=1}^{I}\frac{1}{J_i}\left\{Var(Y_{i\cdot}) + [E(Y_{i\cdot})]^2\right\} - \frac{1}{N}\left\{Var(Y_{\cdot\cdot}) + [E(Y_{\cdot\cdot})]^2\right\}$$

$$= \sum_{i=1}^{I}\frac{1}{J_i}\left\{J_i\sigma^2 + J_i^2(\mu + a_i)^2\right\} - \frac{1}{N}\left\{N\sigma^2 + N^2\mu^2\right\} =$$

$$= (I-1)\sigma^2 + \sum_{i=1}^{I} J_i a_i^2$$

Συνεπώς $E(MStr) = E\left(\frac{SStr}{I-1}\right) = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} J_i a_i^2$.

Στατιστικό τεστ για έλεγχο $H_0: a_1 = a_2 = \ldots = a_I$.

Πρακτική αξία:

Δίνει τη δυνατότητα να διερευνήσουμε αν κάποιο από τα επίπεδα του παράγοντα ασκούν σημαντικότερη επίδραση στην εξαρτημένη μεταβλητή Y.

• Αν η H_0 δεν μπορεί να απορριφθεί, αυτό σημαίνει ότι δεν υπάρχει επίπεδο που να ασκεί σημαντικότερη επίδραση στην Y και η περετέρω διερεύνηση σταματά.

• Αν η H_0 απορριφθεί αυτό σημαίνει ότι κάποιο από τα επίπεδα ασκεί σημαντικότερη επίδραση στην Y.
Επομένως γεννιέται το ερώτημα ποιά από τα επίπεδα είναι σημαντικότερα με την έννοια ότι ασκούν σημαντικότερη επίδραση στην Y.

Για τεχνικούς λόγους θεωρώ την ισοδύναμη υπόθεση: $a_1 = a_2 = \ldots = a_I = 0$

Αποδείξαμε ότι $E(MSres) = \sigma^2$, $E(MStr) = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} J_i a_i^2$

Άρα αν η υπόθεση H_0 είναι αληθής τότε
$E(MSres) = E(MStr)$ ή $MSres \approx MStr$.
Αλλά αν ~~E~~ $MSres \not\approx MStr$ τότε απορρίπτω την H_0.
Επομένως το τεστ θα στηριχτεί στη σύγκριση των MSres και MStr. Προτιμάμε το πηλίκο $\frac{MStr}{MSres}$ διότι εικάζουμε ότι μπορεί να έχει μια F-κατανομή.

Θεώρημα:
Υπό τις υποθέσεις για τα σφάλματα και την $H_0 : a_1 = \ldots = a_I = 0$ ισχύουν τα εξής:

① $\boxed{\frac{SSres}{\sigma^2} \sim X^2_{N-I}}$ ② $\boxed{\frac{SStr}{\sigma^2} \sim X^2_{I-1}}$

Απόδειξη:

① $SSres = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i.})^2$

$Y_{i1}, \ldots Y_{iJ_i}$ είναι τ.δ. από $N(\mu + a_i, \sigma^2)$.
Υπό την H_0 $Y_{i1}, \ldots Y_{iJ_i} \sim N(\mu, \sigma^2)$ τότε
$\frac{(n-1) S_w^2}{\sigma^2} \sim X^2_{n-1}$ Συνεπώς για το $Y_{i1}, \ldots Y_{iJ_i}$ έχουμε

$$\frac{\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2}{\sigma^2} \sim \chi^2_{J_i-1} \Rightarrow \frac{\sum_{i=1}^{I}\sum^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I}(J_i-1)} \Rightarrow$$

$$\Rightarrow \frac{SS_{res}}{\sigma^2} \sim \chi^2_{N-1}$$

② $SS_{tr} = \sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$

Γνωρίζουμε ότι αν $W_1, \dots, W_n$ τ.δ. από $N(\mu, \sigma^2)$ τότε $\bar{W} \sim N(\mu, \frac{\sigma^2}{n})$

κατά συνέπεια $\bar{Y}_{i\cdot} \sim N(\mu, \frac{\sigma^2}{J_i}) \Rightarrow \frac{\bar{Y}_{i\cdot}-\mu}{\sigma/\sqrt{J_i}} \sim N(0,1)$

$$\Rightarrow \frac{\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot}-\mu)^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I}1} = \chi^2_I \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2 \sim \chi^2_{I-1} \Rightarrow \frac{SS_{tr}}{\sigma^2} \sim \chi^2_{I-1}$$

• Τα SS_{res} και SS_{tr} είναι μεταξύ τους ανεξάρτητα.

Για τον έλεγχο της H_0 θεωρώ το F-πηλίκο

$$F = \frac{MS_{tr}}{MS_{res}} = \frac{SS_{tr}/(I-1)}{SS_{res}/(N-1)} = \frac{SS_{tr}/\sigma^2(I-1)}{SS_{res}/\sigma^2(N-1)} \sim \frac{\chi^2_{I-1}/(I-1)}{\chi^2_{N-1}/(N-1)}$$

+ ανεξαρτησία $\Rightarrow F_{I-1,N-1}$ υπό την H_0

Η H_0 απορ. όταν MS_{tr} πολύ διαφορετικό του MS_{res} ή

η H_0 απορ όταν $MS_{tr} \gg MS_{ress}$.

Μορφή κρ. περιοχής $F \geq c$.

Προσδιορισμός κρίσιμου σημείου c.

$a = P(\text{Απορ } H_0 \mid H_0 \text{ αληθής}) = P(F \geq c \mid F \sim F_{I-1,N-1}) =$

$= P(F_{I-1,N-1} \geq c) \Rightarrow \boxed{c = F_{I-1,N-1,a}}$

Συγκεντρωτικά:

Για τον έλεγχο της $H_0: a_1 = \dots = a_I$ (δηλ. του έλεγχο της ισοεπίδρασης των επιπέδων του παράγοντα στην εξαρτημένη μεταβλητή Y) η σ.σ.τ είναι το F πηλίκο $F = \frac{MStr}{MSres}$ με κατανομή την $F_{I-1, N-1}$ υπό την H_0 και κρίσιμη περιοχή μεγέθους α $F \geq C = F_{I-1, N-1, a}$.

Συνήθως η $H_0: a_1 = a_2 = \dots = a_I$ απορρίπτεται. Αυτό σημαίνει ότι κάποιο ή κάποια από τα επίπεδα του παράγοντα ασκούν πιο σημαντική επίδραση στην Y.

Ερώτημα:

Ποιο ή Ποια επίπεδα ασκούν πιο σημαντική επίδραση στην Y.

→ οι πολλαπλές συγκρίσεις δίνουν την απάντηση διότι συγκρίνουν τα επίπεδα και τα κατατάσσουν σε κατηγορίες σημαντικότητας

Μέθοδοι πολλαπλών συγκρίσεων (δηλ. συγκρίσεων των επιπέδων του παράγοντα ανά δύο)

1) Ελάχιστη σημαντική διαφορά
2) Tukey
3) Bonferoni
4) Γραμμικών αντιθέσεων ή Scheffe.

Μέθοδος της Ελάχιστης Σημαντικής Διαφοράς (ΕΣΔ).

Αν η γενική υπόθεση της ισοεπίδρασης $H_0: a_1 = a_2 = \dots = a_I$ απορριφθεί τότε καταφεύγουμε σε συγκρίσεις ανά δύο. Δηλαδή ελέγχουμε τις υποθέσεις.

$H_0: a_i = a_{i'}$ v $H_a: a_i \neq a_{i'}$, $\forall i, i' = 1, 2, \dots I$.

Για τον έλεγχο της $H_0: a_i = a_{i'}$ στηριζόμαστε σε έναν εκτιμητή των παραμέτρων που εμφανίζονται στην H_0 (η εκτίμηση κατά Wald θεώρηση)

$$\left.\begin{aligned} \hat{a}_i &= \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \\ \hat{a}_{i'} &= \bar{Y}_{i'\cdot} - \bar{Y}_{\cdot\cdot} \end{aligned}\right\} \Rightarrow \hat{a}_i - \hat{a}_{i'} = \bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}$$

Θα στηριχτώ στην διαφορά $\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}$ και αρκεί να βρω την κατανομή της υπό την $H_0: a_i = a_{i'}$.

$Y_{i1}, \dots, Y_{iJ_i}$ τ.δ από κατανομή με $N(\mu + a_i, \sigma^2)$ τότε

$$\bar{Y}_{i\cdot} \sim N\left(\mu + a_i, \frac{\sigma^2}{J_i}\right) \quad (1)$$

Επίσης $Y_{i'1}, \dots, Y_{i'J_{i'}}$ τ.δ από $N(\mu + a_{i'}, \sigma^2)$ τότε

$$\bar{Y}_{i'\cdot} \sim N\left(\mu + a_{i'}, \frac{\sigma^2}{J_{i'}}\right) \quad (2)$$

Τότε από (1), (2) + ανεξαρτησία $\bar{Y}_{i\cdot}, \bar{Y}_{i'\cdot}$

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \sim \left(0, \frac{\sigma^2}{J_i} + \frac{\sigma^2}{J_{i'}}\right)$$ υπό την $H_0: a_i = a_{i'}$

$$\left.\begin{array}{l} \Rightarrow \dfrac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sigma\sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}} \sim N(0,1) \\[2ex] \dfrac{SS_{res}}{\sigma^2} \sim \chi^2_{N-1} \end{array}\right\} \text{Διαιρώ}$$

$$\Rightarrow \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{MS_{res}} \cdot \sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}} \sim t_{N-I}$$ υπό την H_0.

Για τον έλεγχο λοιπόν της $H_0: a_i = a_{i'}$ η σ.σ.τ. είναι η

$$t = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{MS_{res}} \cdot \sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}} \sim t_{N-1}$$

και κρ. περιοχή μεγάλες περιοχές του πηλίκου:

$$\left| \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{MS_{res}} \sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}} \right| \geq c = \dots = t_{N-1, \frac{a}{2}}.$$

ή $|\overline{Y_{i\cdot}} - \overline{Y_{i'\cdot}}| \geq \underbrace{t_{N-I, \frac{a}{2}} \sqrt{MSres} \sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}}$

υπολογίσιμη ποσότητα που ονομάζεται ελάχιστη σημαντική διαφορά

Εφαρμογή στην πράξη:
Η μέθοδος της ΕΣΔ εφαρμόζεται ως εξής:

Αν $|\overline{Y_{i\cdot}} - \overline{Y_{i'\cdot}}| > (ESD)$ τότε η H_0: απορ.
Αν $H_0: a_i = a_{i'}$ απορ. τότε κάποιο από τα επίπεδα i, i' ασκεί πιο σημαντική επίδραση στην Y.
Για τον έλεγχο για τον ποιο ασκεί σημαντικότερη επίδραση απο τα i και i' εκτελώ το εξής:

- Αν $\overline{Y_{i\cdot}} - \overline{Y_{i'\cdot}} > 0 \Rightarrow$ i-επίπεδο ασκεί πιο σημαντική επίδραση από το i'
- Αν $\overline{Y_{i\cdot}} - \overline{Y_{i'\cdot}} < 0 \Rightarrow$ i' επίπεδο ασκεί πιο σημαντική επίδραση απο το i

Ασκήσεις

1/ Στο μοντέλο της ΑΝΑΔΙΑ κατα έναν παράγοντα ν.δ.ο οι εκτιμητές ελαχίστων τετραγ. των παραμέτρων μ και αi είναι αμερόληπτοι (δηλ $E(\hat{\mu}) = \mu$, $E(\hat{a}_i) = a_i$, $i = 1, \ldots I$)

(2) Στο μοντέλο ΑΝΑΔΙΑ κατα έναν παράγοντα ν.δ.ο οι εκτιμητές ελ. τετραγώνων του μοντέλου συμπίπτουν με τους Ε.Μ.Π.